DISS. ETH. NO. 27561

Transcriptional recording by CRISPR spacer acquisition from RNA

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

FLORIAN SCHMIDT

M.Sc. Molecular Biotechnology, Ruprecht Karl University of Heidelberg

born on 03.09.1991

citizen of Germany

accepted on the recommendation of

Prof. Dr. Randall J. Platt (ETH Zurich, Switzerland), examiner

Prof. Dr. Yaakov Benenson (ETH Zurich, Switzerland), co-examiner

Prof. Dr. Michael P. Terns (University of Georgia, United States of America), co-examiner

2021

Abstract

Our capacity to investigate the cellular transcriptome is crucial to interrogate complex cellular behaviors and states and therefore fundamental for understanding cellular life.
The advent of deep-sequencing technologies, in particular RNA sequencing (RNA-seq) has endowed researchers with the ability to investigate the cellular transcriptome at unprecedented depth, thereby shedding light on complex cellular states. However, RNA-seq based technologies are disruptive by nature limiting observations to a single moment in time or the collection of time-resolved data from asynchronous samples.

Molecular recording technologies based on DNA modifying enzymes can overcome this limitation by encoding cellular stimuli into DNA, thereby preserving information over time. However, current molecular recording technologies are intimately tied and therefore limited to appropriate biosensors and accordingly restricted to recording defined stimuli, narrowing their application to a small set of molecules and cellular pathways.

The microbial adaptive immune system CRISPR (clustered regularly interspaced short palindromic repeat) encodes the CRISPR associated (Cas) proteins Cas1 and Cas2 which together form the CRISPR adaptation complex (Cas1–Cas2). This protein machinery constitutes the core of adaptive immunity by sampling cellular DNA and incorporating small snippets of DNA into a genomic CRISPR array. Thereby Cas1–Cas2 record creates a molecular fossil record of invading genetic elements that is inherited from one generation of cells to their daughters. A recently discovered variant of this adaptation complex contains a natural fusion of a reverse transcriptase (RT) to Cas1, resulting in an RT-Cas1–Cas2 complex that can capture and convert intracellular RNAs into DNA when expressed in its host bacterium but is deficient of RNA spacer acquisition when employed in an *E. coli* host.

In this thesis, we discover and characterize an RT-Cas1–Cas2 ortholog from the human commensal bacterium *Fusicatenibacter saccharivorans* - *Fs*RT-Cas1–Cas2 - enabling transcriptional recording in *E. coli* and thereby facilitating storage of transcriptome-scale information into plasmid DNA. We recover transcriptional records by developing selective amplification of expanded CRISPR arrays (SENECA) followed by deep-sequencing. Aligning the acquired CRISPR spacers to the *E. coli* genome enables transcript quantification and associated differential expression analysis. Using this Record-seq workflow *in vitro* we can

record both simple and complex transcriptional events such as RNA virus infection, oxidative stress and exposure to herbicides, pinpointing the precise genes involved in cellular responses. We leverage *E. coli* cells equipped with *Fs*RT-Cas1–Cas2 as sentinels traversing and monitoring the murine gastrointestinal tract while recording their environment. This allows us to assess physiological and pathophysiological states of the hosts intestine in a non-invasive fashion. The application of Record-seq derived sentinel cells in the gut allows us to investigate a broad range of manipulations *in vivo* such as alterations from the host diet, chemically induced inflammation of the intestine and the interaction of our sentinel cells with host microbiota. These applications broadly establish Record-seq as a non-invasive technology for transcriptional recording to understand biological processes in health and disease.

Zusammenfassung

Unsere Fähigkeit das zelluläre Transkriptom zu erforschen ist essentiell für die Erforschung von komplexen zellulären Verhaltensweisen und daher grundlegend zum Verständnis des Lebens.

Mit dem Aufkommen der Hochdurchsatzsequenzierung, insbesondere der Technologie der RNA-Sequenzierung (RNA-seq) ist es Wissenschaftlern ermöglicht worden, das zelluläre Transkriptom in nie gekannter Tiefe zu untersuchen und hiermit komplexe Zellstadien zu verstehen. Allerdings beruhen RNA-seq Technologien stets auf Methoden, welche die Zelle zerstören und sind daher auf einen einzelnen Zeitpunkt oder die Untersuchung von Ansammlung von zeitaufgelösten Daten von asynchronen Proben limitiert.

Molekulare Aufnahmegeräte basierend auf DNA modifizierenden Enzymen können solche Limitationen überwinden, indem sie zelluläre Signale in DNA überschreiben und somit die Information über lange Zeiten konservieren. Allerdings sind aktuelle molekulare Aufnahmegeräte eng an die Verwendung von passenden Biosensoren geknüpft und daher auch durch diese Biosensoren in ihrer Funktionalität eingeschränkt. Dies restriktiert sie auf die Aufzeichnung von definierten Signalen und limitiert ihre Anwendung auf eine kleine Anzahl von Biomolekülen und zellulären Signalwegen.

Das adaptive mikrobielle Immunsystem CRISPR (aus dem Englischen clustered regularly interspaced short palindromic repeats - gehäuft auftretende, regelmäßig unterbrochene, kurze palindromische Wiederholungen) codiert die CRISPR assozierten (Cas) Enzyme Cas1 und Cas2. Zusammen formen diese den sogenannten adaptiven CRISPR Komplex (Cas1–Cas2). Dieser Proteinkomplex stellt den Grundstein der adaptiven Immunität dar. Der Cas1–Cas2 Komplex nimmt Stichproben der intrazellulären DNA und integriert kleine Fragmente (Spacer) von dieser in eine spezifische Position des mikrobiellen Genoms – das sogenannte CRISPR Array. Dadurch kreiert Cas1–Cas2 eine molekulare, fossile Überlieferung von feindlichen genetischen Elementen welche die Zelle befallen haben. Dieses CRISPR Array wird von einer Zelle an ihre Tochterzellen weitervererbt. Eine kürzlich beschriebe Variante dieses Cas1–Cas2 Komplexes verfügt über ein Cas1 Enzym, welches natürlicherweise an eine Reverse Transkriptase (RT) fusioniert ist. Dies resultiert in einem RT-Cas1–Cas2 Komplex welcher auch intrazelluläre RNA Fragmente einfangen und in DNA überschreiben kann. Allerdings war

dieser neu entdeckte RT-Cas1–Cas2 Komplex nur in seinem Ursprungsbakterium funktional, während er in *E. coli* keine RNA Fragmente mehr als Spacer in DNA überschreiben konnte.

In dieser Doktorarbeit entdeckte und charakterisierte ich einen orthologen RT-Cas1–Cas2 Komplex des symbiontischen menschlichen Bakteriums *Fusicatenibacter saccharivorans* - kurz *Fs*RT-Cas1–Cas2 – welcher auch in *E. coli* in der Lage ist RNA in DNA spacer umzuschreiben und hierdurch als Aufnahmegerät für das zelluläre Transkriptom verwendet werden kann und so die transkriptomweite Speicherung von Informationen in Plasmid-DNA ermöglicht. Durch die Entwicklung einer neuen Methode genannt SENECA (aus dem Englischen selective amplification of expanded CRISPR arrays – selektive Verfielfältigung von erweiterten CRISPR Arrays) wurde es in Kombination mit der Hochdurchsatzsequenzierung möglich diese Transkriptionalen Aufnahmen aus der DNA wiederzuerlangen. Die sequenzierten Spacer werden mit dem Genom von *E. coli* abgeglichen um so das Transkriptom des Bakteriums zu quantifizieren und Genexpressionsanalyse und vergleichbare Methoden durchzuführen – eine Technologie die wir als Record-seq bezeichnen. Indem wir diese Record-seq Methode *in vitro* anwenden ist es uns möglich sowohl einfache als auch komplexe Transkriptionsereignisse nachzuvollziehen. Beispielsweise die Infektion von *E. coli* mit einem RNA phagen, oxidativen Stress oder die Exposition des Bakteriums zu Herbiziden. Hierbei ist es möglich präzise jene Gene zu Identifizieren die an der zellulären Antwort zu dem entsprechenden Ereignis beteiligt waren.

Darüber hinaus haben wir diese mit *Fs*RT-Cas1–Cas2 ausgestatten *E. coli* Zellen auch als sogenannte Wächterzellen im Darm von Mäusen eingesetzt. Die Zellen passieren den Gastrointestinaltrakt gemeinsam mit der Nahrung und zeichnen die dort vorherrschenden Bedingungen auf. Dies ermöglicht es uns den physiologischen und pathophysiologischen Status der Wirtsmaus zu bestimmen ohne hierfür invasive, chirurgische Methoden zu verwenden. Dadurch ermöglichte es Record-seq ein weites Portfolio an Manipulationen und Experimenten *in vivo* zu untersuchen. Beispielsweise verschiedene Variationen in der Nahrungszusammensetzung der Maus, chemisch induzierte Dickdarmentzündung oder aber auch die Interaktion unserer Wächterzellen mit anderen Bakterien der Darmflora.
Zusammengenommen etablieren diese Anwendungsmöglichkeiten Record-seq als eine nichtinvasive Technologie die es ermöglicht das zelluläre Transkriptom aufzunehmen und somit biologische Prozesse im gesunden wie erkrankten Lebewesen zu verstehen.